ADAPTIVE SPATIO-TEMPORAL FILTER FOR LOW-COST CAMERA DEPTH MAPS

Massimo Camplani and Luis Salgado

Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid, Spain

ABSTRACT

In this paper we present an adaptive spatio-temporal filter that aims to improve low-cost depth camera accuracy and stability over time. The proposed system is composed by three blocks that are used to build a reliable depth map of static scenes. An adaptive joint-bilateral filter is used to obtain consistent depth maps by jointly considering depth and video information and by adapting its parameters to different levels of estimated noise. Kalman filters are used to reduce the temporal random fluctuations of the measurements. Finally an interpolation algorithm is used to obtain consistent depth maps in the regions where the depth information is not available. Results show that this approach allows to considerably improve the depth maps quality by considering spatio-temporal information and by adapting its parameters to different levels of noise.

Index Terms— Depth camera, Depth Map denoising, Joint-Bilateral Filter, Hole Filling

1. INTRODUCTION

The availability in the market of cheap depth cameras like [1] is increasing the potentiality of computer vision applications and, at the same time, is capturing the attention of the research community. This new family of devices presents two main advantages with respect to its competitors in the depth cameras market: it allows to obtain dense depth maps with higher resolution and its price is considerably lower than other active depth sensors. The research community of computer vision and the game industry are developing several *depth camera based* applications mainly related to human-computer interaction. Human silhouette extraction and skeleton tracking are the most investigated topics, especially for controller-free video games [2]. Other examples of low-cost depth camera applications are [3] where it is used as a touch sensor, or [4] where it is used in a multi-media learning environment. Low-cost depth cameras have the potentiality to be widely employed in several domains, not only in the game industry, but also in other applications such as immersive and interactive environments, SLAM systems, surveillance applications etc.. However, these devices present noise-related problems that have to be solved to improve the quality of the depth information and to broaden their application possibilities.

Depth maps generated by active or passive cameras are generally characterized by a high level of noise and, for this reason, several techniques have been developed to improve their accuracy. These approaches aim at refining depth discontinuities at object edges, obtaining a smooth depth map for regions belonging to the same object and therefore removing artifacts. Gaussian filtering and other smoothing approaches lead to poor results especially in depth discontinuities where the map appears obviously blurred; for this reason, in literature, many approaches based on edge preserving techniques are deployed to avoid the blurring effect. One of the most used edge-preserving filtering technique is the bilateral filter presented in [5], where the filter weights are selected as a function of a photometric similarity measure of the neighbor pixels. In this approach the edges are preserved since non-similar neighbor pixels are not considered in the filtering process. This idea has been extended in the joint-bilateral or cross-bilateral filter presented in [6], where the filter weights are selected as a function of the properties of another image. This approach has been widely used for depth map denoising, since visual information can be used to improve the accuracy of the depth map near object contours [7, 8], it has been employed to reduce the noise of depth images that is generated by an upsampling procedure; in [9], a tri-lateral filter is used to reduce the noise in depth maps obtained with stereo pairs, considering also a confidence measure of the stereo matching algorithm in the filter weights selection; in [10], an adaptive joint-bilater filter is iteratively applied to refine the depth map.

The objective of this paper is to highlight the main problems of the low-cost depth camera model presented in [1] and to propose an efficient solution to address them. In particular, we present an adaptive spatio-temporal filtering approach that aims to build a reliable and consistent depth map. The system is composed by three main blocks: an adaptive joint-bilateral filter (AJBF) is used to correct the depth map by integrating depth and video information and by adapting its parameters to the noise level; an adaptive Kalman filter (AKF) is used for each pixel to reduce random fluctuation of its value and to estimate a reliable depth map model of the scene; finally, an

This work has been supported by the Ministerio de Ciencia e Innovaciòn of the Spanish Gov. under proj. TEC2010-20412 (Enhanced 3DTV). Massimo Camplani would like to acknowledge the European Union and the Universidad Politécnica de Madrid for supporting his reserach activities through the Marie Curie-Cofund research grant.

efficient interpolation system (HF) is used to fill the areas for which the depth measurements are missing.

2. DEPTH MAP ISSUES

The low cost depth camera presented in [1] (that from now on will be called DC) is a structured light 3D scanner, where an infrared light source projects light patterns to the space: the reflected light is received by an infrared camera and processed to extract geometric information about object surfaces. Structured light sensors measurements are affected by noise due to multiple reflections, transparent objects or scattering in particular surfaces (i.e. human tissues), as it happens for other active optical sensors (i.e. tof cameras [11]). The main objective of this section is to highlight the noise effects in the acquired DC depth maps. For the best of authors' knowledge only the work in [12] present a detailed study on the precision of the DC that supports the results obtained in our tests. In

Fig. 1. Video data(a) and corresponding depth map (b)

Figure 1 the visual image (left) and the depth map (right) of an indoor environment obtained with the DC are shown. As it can be noticed, for some pixels of the depth map (marked in red), the camera is not able to provide depth information. These *no-measured depth* pixels (*nmd* pixels) appears mainly in correspondence of concave objects (such as empty spaces in the library), but *nmd* pixels are also present in homogeneous regions of the images (i.e. above the door); moreover *nmd* pixels are present near object boundaries in presence of high depth discontinuities (i.e the area close to the box), therefore an efficient strategy to fill these areas is required.

Figure 1 highlights also that the DC depth maps are affected by a high level of noise in the object boundaries regions that are very irregular and, thus, not accurate. On the contrary the edges that can been extracted from the video data are more robust, for this reason the proposed filtering strategy has to take into account also the visual information to improve the depth map accuracy.

Another error that affects the DC measurements is the single pixel stability over time. In fact, the depth measurements for a single pixel related to a fixed object vary along time and therefore, a temporal smoothing approach is needed to reduce these random fluctuations. As we can observe in Figure 1(wall and door regions), another error in the depth map is that neighbor pixels corresponding to object positioned at the same real distance with respect to the camera present different depth values. A quantitative example of this measurement error is shown in Figure 2 where the frequency histogram of the estimated depth of an orthogonal flat region (positioned at 3.6 meters with respect to the camera) is shown. This example and other tests demonstrate that the error introduced in depth measurements of orthogonal flat regions can be modeled with a Gaussian distribution; in this case, the Gaussian distribution (red line) has a mean value corresponding to 3.61 meters, that is a good estimation of the real distance.

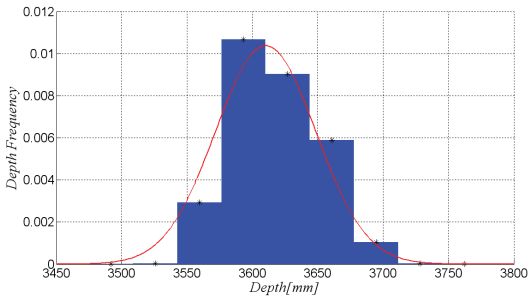

Fig. 2. Distribution of measured depth in a neighborhood.

In order to analyze the impact of this error at different distances, tests have been conducted by positioning the camera orthogonally to a white wall and varying its distance from it (from 0.86 cm up to 3.86 m) and considering a set of 200 frames. In Fig 3 we report the variance σ of the measured depth as a function of the wall-camera distance (solid line): it is worth noting that their relationship could be modeled as a linear function (dashed line) and that σ increases with the distance. This analysis shows that the depth measurements are affected by a noise that depends on the camera-object distance, therefore an efficient noise removal strategy that consider also this factor is required.

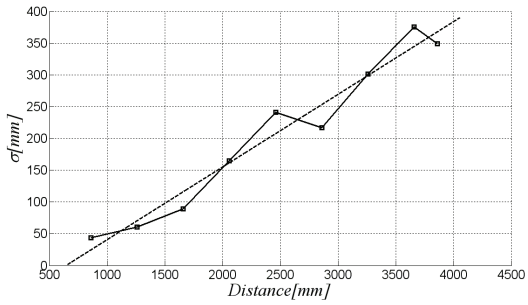

Fig. 3. σ as a function of the depth (black line) and σ-depth linear regression (dashed line)

In this section we have presented an overview of the most important problems that affect the accuracy of DC-generated depth maps: *nmd* pixels, noisy object boundaries,

depth measurements fluctuation, incoherent neighborhoods and distance dependent noise. Therefore, it appears necessary to investigate techniques to efficiently tackle these problems and, in this way, improving the accuracy of DC-generated depth maps. In the following section, the strategies that we propose to address these issues are presented.

3. ADAPTIVE SPATIO-TEMPORAL FILTER

The proposed spatio-temporal filter is composed of three different blocks, as shown in Figure 4: the *Adaptive Joint-Bilateral Filter* (AJBF) used to reduced the spatial noise and, at the same time, to improve the depth map near objects boundaries; the *Hole Filling* (HF) block is in charge of completing the regions of the depth image where *nmd* pixels are present; an *Adaptive Kalman Filter* (AKF) approach used to keep track of the measurements temporal fluctuation. The main idea of our approach is to continuously improve a depth map model (D_{mod}) by applying an adaptive temporal smoothing with the AKF. The depth model is recursively used by the AJBF to smooth the image homogeneous regions and to correct the object contours. The AJBF considers also the the intensity image (I) provided by the DC video sensor. HF corrects the holes in the depth map, and the result of this processing is used to update the model.

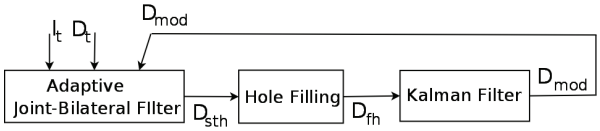

Fig. 4. Main blocks of the proposed approach

The AJBF reduces the spatial variation of depth values that belongs to the same object region while contemporary preserving and improving the edges depth measurements. Our approach is based on bilateral filter [5] and its extension joint-bilateral filter [6]. In particular, the presented AJBF has been designed in order to filter the depth map by considering the information of the intensity image and the information of the depth model that is iteratively estimated by the system.

Let us consider D_p the depth value of the pixel at position p and Ω_p its neighborhood, the resulting filtered pixel J_p obtained by the AJBF is

$$J_p = \frac{1}{k_p} \sum_{q \in \Omega_p} D_q f(p,q) g(\left\| D_p^{mod} - D_q^{mod} \right\|) h(\|I_p - I_q\|) \quad (1)$$

where $f(p,q)$ is a 2D smoothing kernel also known as *domain* term, that measures the closeness of the pixels; $g(\left\| D_p^{mod} - D_q^{mod} \right\|)$ is the *range* term that measures the pixels similarity of the modeled depth map ; $h(\|I_p - I_q\|)$ is the range term that measures the similarity in the intensity domain . The scalar k_p is a normalization factor. The functions f,g and h are modeled as Gaussian. The proposed Joint-Bilateral filter is adaptive since the gaussian function used for $g(x)$ is characterized by a variable σ_r (variance of the range filter) that is modified according to the depth value D_p following the estimated profile presented in Figure 3. In this way, the influence of different levels of noise to the depth measurements is considered in our filter model: for high (small) values of measured depths, the value of σ_r is increased (decreased) allowing to properly reduce the noise effect.

The HF block processes the filtered depth map provided by AJBF in particular a 2D gaussian Kernel (similar to $f()$ in equation 1) is used to obtain consistent values for *nmd* pixels. HF includes in the filtering process only those pixels of the neighborhood Ω that present a stable depth value in the model. In fact, with the depth model is also continuously updated a frequency map that give a score about pixels depth value stability. In this way, the *nmd* are corrected iteratively and the reliability of their depth values increases while new samples of that regions are acquired and included in the model.

The AKF is used to complete the depth model: for each pixel it reduces the temporal fluctuations of the pixel values and, consequently, it improves the accuracy of the depth map model. The Kalman Filter system is adaptive since, for different depths, different models of noise (with different variances) are used as in the case of the AJBF. As previously mentioned, the generated depth map model, is used twice by the adaptive joint-bilateral filter: once to estimate the new parameters σ_r and then to calculate the filter weights.

4. RESULTS

In section 2 the main problems of the DC depth map have been presented. The resulting depth map obtained with the proposed approach is presented in Figure 5. As it can be noticed, the errors from the depth map have been successfully reduced: flat regions present in this case a uniform value without any *nmd*. The edges of the objects have been accurately refined also in the concave zone of the library. However, the hole filling approach for those areas is clearly less efficient than in the case of flat regions such as in the upper part of the wall. In Figure 6 a detail of the video frame and the results obtained applying different filtering approaches are shown. In particular, with the proposed method (Figure 6 (c)) we obtain a much more accurate depth map, especially near the object boundaries, than the one obtained with a bilateral filter that considers only the depth information (Figure 6 (b)). Moreover, the advantages of the iterative filtering process (through AKF) are shown in Figure 7 where depth map model estimated after the processing of one frame (Figure 7 (b)) and ten frames (Figure 7 (c)) are shown: depth map accuracy is improved with the temporal smoothing process.

Fig. 5. Depth map obtained with the proposed method.

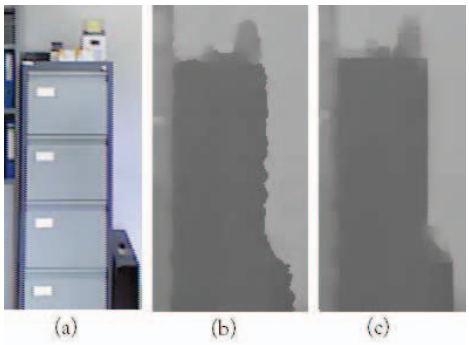

Fig. 6. Video data(a), bilateral filter (b), AJBF (c)

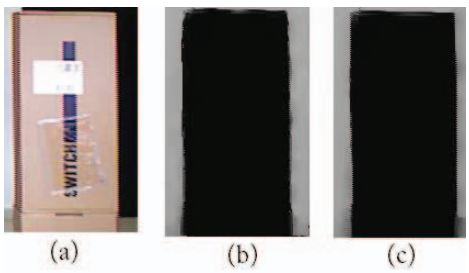

Fig. 7. Video data(a) and corresponding depth map model after one (b) and ten frames(c)

5. CONCLUSION

In this paper an adaptive filtering system for a low cost depth camera map has been presented. Low-cost depth camera presents several problems that affect the accuracy of the depth maps: *nmd* pixels, depth measurements fluctuations, incoherent neighborhoods, distance dependent noise and noisy edges. The proposed system allows to successfully address these issues: an adaptive joint-bilateral filter, based on the analysis of the visual and depth information, is used to reduce the spatial variance of the measured depth without introducing any blurring effect on the edges. Moreover the adaptive filter modifies its parameters to different level of noise. An efficient hole filling procedure is used to iteratively improve the depth map model. Finally, an Adaptive Kalman Filter is employed to reduce the random fluctuations of measured depth values over time and to build a consistent model of the depth map. The same model is iteratively improved and used to accurately adapt the AJBF parameters. The results show that the proposed approach is able to efficiently tackle all the problems related to depth maps generated by the depth camera.

6. REFERENCES

[1] "The xbox kinect," `http://www.xbox.com/kinect`.

[2] Amit Bleiweiss et al., "Enhanced interactive gaming by blending full-body tracking and gesture animation," in *ACM SIGGRAPH ASIA 2010 Sketches*, 2010, SA '10, pp. 34:1–34:2.

[3] Andrew D. Wilson, "Using a depth camera as a touch sensor," in *ACM International Conference on Interactive Tabletops and Surfaces*, 2010, ITS '10, pp. 69–72.

[4] M. Roccetti et al., "The art and craft of making the tortellino: playing with a digital gesture recognizer for preparing pasta culinary recipes," *Comput. Entertain.*, vol. 8, pp. 28:1–28:20, December 2010.

[5] C Tomasi and R Manduchi, "Bilateral filtering for gray and color images," *Sixth International Conference on Computer Vision*, vol. 846, pp. 839–846, 1998.

[6] Georg Petschnigg et al., "Digital photography with flash and no-flash image pairs," in *ACM SIGGRAPH 2004 Papers*, 2004, pp. 664–672.

[7] O.P. Gangwal and B. Djapic, "Real-time implementation of depth map post-processing for 3d-tv in dedicated hardware," in *International Conference on Consumer Electronics (ICCE)*, jan. 2010, pp. 173 –174.

[8] Johannes Kopf et al., "Joint bilateral upsampling," in *ACM SIGGRAPH 2007 papers*. 2007, ACM.

[9] M. Mueller et al., "Adaptive cross-trilateral depth map filtering," in *3DTV-Conference*, 2010, pp. 1 –4.

[10] PoLin Lai et al., "Depth map processing with iterative joint multilateral filtering," in *Picture Coding Symposium (PCS), 2010*, dec. 2010, pp. 9 –12.

[11] S. Foix et al., "Lock-in time-of-flight (tof) cameras: A survey," *Sensors Journal, IEEE*, , no. 99, pp. 1, 2011.

[12] Fabio Menna et al., "Geometric investigation of a gaming active device," *SPIE*, vol. 8085, no. 1, pp. 80850G, 2011.